

---

# Multihop Multimodal QA on WebQA Dataset

---

Sushil Khyalia<sup>1</sup> Shruti Bansal<sup>1</sup> Manh-Bao Nguyen<sup>1</sup> Avi<sup>1</sup>

## 1. Abstract

In this paper, we explore the problem of multimodal multihop QA on the WebQA dataset. This problem statement is highly relevant to the fast evolving world of virtual assistants and chatbots, expected to retrieve relevant multimodal content available to answer the user’s query. Latest approaches are still underperformant on simple questions for humans, often overlooking the visual hints. Besides, an efficient unified pipeline for retrieval and generation based on transparent multi-hop reasoning still seems far away. Our research ideas tackle different aspects of these challenges: in our first idea, we explore sequential retrieval with universal embeddings from verbalized images, which we also use for generation in a unified pipeline. For our second, we propose the use of reformulation of questions by decomposing them into sub-questions and abridging them to generate questions which promote multi-hop answering. Finally, we optimize for synergy maximisation by changing the loss function while training on existing baselines. All these explorations give encouraging ideas for more transparent human-like multihop reasoning for question answering, which avoids relying on biases by leveraging all relevant sources, in a more end-to-end framework.

## 2. Introduction

Multi-hop multi-modal QA mimics the way humans perceive and process information. We do not restrict ourselves to one source of information, but aggregate the information from multiple sources around us, whether those are visual clues or text or audio, in line with the concept of multimodality. More often than not, we process information in multiple steps, making deductions and associations as we try to find the answer to a question. Multi-hop architectures are trying to induce a similar behavior in machines in the domain of visual question answering. For making IOT devices or virtual assistants more adept at deriving useful information from the web or other sources of information which include multiple modalities, they need to learn deductive reasoning based on multiple sources. It is the natural step in the progression from human-initiated searches on the internet to fully autonomous agents capable of extracting information from all resources available and aggregating the information.

Indeed, as of now, humans naturally integrate multi-modal information and abstract multi-hop links over them so as to find information. On the internet, humans can not easily process aggregate all the different sources and instead adapt by breaking their complex question into single-hop sub-questions, each answered by a simple precise query and associated reasoning. In the Machine Learning community, the questions of multimodal question-answering (eg: VQA) and multi-hop textual question-answering have been studied for a long time, and some recent works tackle this intersection of multi-modal and multi-hop challenges. The latter are of great interest not only for the general public (who already uses ChatGPT to help them with general question-answering) but also for the scientific community to better understand multimodal interactions for reasoning.

In this paper, we propose a research idea based on the decomposition of the Questions present in the dataset into simpler subquestions which can be used to implement a multi-hop architecture for the multihop questions in the WebQA dataset that require more than once source to find the answer. In the following sections, we explain the related works that have worked on this line of thought before. In the proposed approach we discuss the novel multi-hop multimodal architecture for answering Intersection questions in WebQA and similar datasets. In the experimental setup, we give detailed explanation of the procedure and the various experiments performed using different models. Another avenue of research we explore is whether the optimization problem for question-answering can be formulated such that we force the model to use both sources and query to get the final answer instead of over-relying on either one of them. We explore one such change to the optimization problem and discuss the issues with that formulation and why it doesn’t improve performance for WebQA. Finally we try to come up with single pipeline for both retrieval and generation using Universal CLIP based embeddings fine tuned for retrieval in a multi-hop manner and use the same embeddings for generation.

We will use this Github repository for the project: <https://github.com/awi121/11-777-Project>

### 3. Related Works

#### 3.1. Multimodal Input to LLMs

Multi-image question answering requires reasoning over multiple sources which is inherent in Large text question answering models. Over the past few years GPT based models have shown State of the art results for question answering over in large scale text comprehension hence they do have multi-hop capabilities. (Alayrac et al., 2022) introduced large scale image and text generative models which utilized both LLMs and Visual Encoders for training LLMs with additional layers. Flamingo has atleast 3B trainable parameters along with frozen image encoders and LLM. (Liu et al., 2023a) also maps images to LLM output with instruction tuning utilizing vision encoders and LLM. This also involves large scale training of huge parameters. (Koh et al., 2023) and (Zhu et al., 2023) work on similar principles but have frozen parameters for both the LLM and Vision encoder and only train a linear layer to ground the image to language embedding space. Using these ideas we we train a small network to convert object detection features and CLIP based retrieval embeddings for multimodal multihop question answering.

#### 3.2. Multimodal Question Decomposition Model

In (Mavi et al., 2022), the authors talk about the need for understanding the multi-step reasoning being performed by models in the background to answer multi-hop questions. To implement this, we need to be able to decompose complex questions into sub-questions that can be answered by the model based on the retrieved information. One paper which explores this idea for text-only models is (Sun et al., 2021), it proposes a new model called DocHopper to attend to different parts of a long documents to answer complex questions. It uses a query  $q$  to derive information from a document and then combined this retrieved information with  $q$  to produce the next query in embedded space. We can extend this idea to multimodal approach for multi-hop QA aggregating information retrieved at each hop in a final hop. This would potentially help in improving the interactions between the modalities. (Chen et al., 2021) also proposes a method to extract a series of sentences as a reasoning chain leading to the answer and is used for predicting the final answer by using BERT. The paper on (Malon & Bai, 2020) talks about generating follow up questions for Multi-hop questions with respect to the HotpotQA dataset. It gives the concept of a neural question generation network for text generation of follow up questions. In (Min et al., 2019), the authors propose to decompose complex multi-modal questions into sub-questions which can be answered individually and the answers combined. Our method extends the idea of question decomposition to multimodal domain by using the decomposed questions on multimodal QA models for both

text and image based queries and using a text only decoder model to combine the answers obtained from each of these multimodal QA models.

#### 3.3. Towards aligned embeddings for retrieval and generation

##### 3.3.1. BRIDGING THE MODALITY GAP

To make the retrieval scalable while linking fine-grained multimodal information, (Liu et al., 2023b) extend contrastive learning for the specific task of source selection, using image verbalization (i.e. automatic enhanced captioning) to bridge the modality gap between image and text. It extends the idea proposed in (Xiong et al., 2021) to multimodal context where all multimodal resources are in the same embedding space. (Yu et al., 2023) goes a step further and proposes to verbalize all modalities to text so as to perform dense retrieval and generation from natural language, leveraging LLMs. However, by doing so, we lose some visual information that cannot be conveyed in a limited span. (Liu et al., 2023b) avoids this pitfall by representing the image by the sum of its visual raw embedding and its verbalized caption.

##### 3.3.2. TOWARDS A MULTI-HOP RETRIEVAL

The retrieval and generation are closely linked by the multi-hop aspect of reasoning. This student report examines retrieval with graph neural networks with different structures. This is intuitively a good start to approach the multi-hop aspect of retrieval, but this performs worse than the baselines in (Chang et al., 2022) which consider each source independently. A possible explanation is that the embedding of the sources were not aligned enough. Mixing fast retrieval over clever dense representations with a progressive retrieval process is explored in the recent paper (Yang et al., 2023b). Contrary to (Liu et al., 2023b), they use different embeddings for the query and the text sources. Besides, the embeddings used for the retrieval are not used for the generation. However, we can take up their idea of deriving a multihop-retrieval objective for their binary classifier.

### 4. Proposed Approach

#### Dataset and Input Modalities

The WebQA dataset is composed of 41732 questions with the associated candidate sources and expected answer from language and vision modalities. There are 20267 text-based questions (the answer is inferred from one or more text based evidences) and 21465 image-based questions (answer is derived from one or more images based evidences). For each query, the dataset provides:

- A question.

- An expected answer.
- A topic.
- The split: train or val.
- A question category: *text* for text-based queries, while the image-based queries are broken down into the following categories for result analysis: *color*, *shape*, *yes/no*, *number*, *choose or others*.
- A set of negative sources, which serve as distractors for the retrieval part: every query comes with both image (image and caption) and text (snippet) distractors.
- A set of positive sources, which are to be retrieved and used to generate the answer. There is a dichotomy here: questions are either related to some images (and their captions), or to some text snippets only.

The dataset comes with train, validation, and test splits and has 36766, 7540, and 4966 examples in each split respectively.

#### 4.1. Universal retrieval and generation using Univl-DR

##### 4.1.1. OVERALL APPROACH AND CONTRIBUTIONS

As mentioned in 3.3.1, (Liu et al., 2023b) leverage CLIP-based dual encoders for the task of source selection after image verbalization. The retrieval is now scalable, and the metrics in the full-retrieval setting are encouraging, showing the potential of image verbalization and integrating both modalities to embed visual sources. Building on that, we want to address certain concerns on the actual retrieval binary choice for each potential source and investigate using such embeddings for generation, paving the way for a scalable and efficient end-to-end QA pipeline. Our ideas of improvements are the following:

- Use better image verbalization methods. Enhancing the initial image captions for further modality alignment is promising, but the current MLM objective in (Liu et al., 2023b) yields very short uninformative captions. (Yang et al., 2023b) and (Yu et al., 2023) use two verbalization models, providing one general description of the image as well as a sequence of local details spotted in the image.
- So far, the contrastive loss in (Liu et al., 2023b) focuses on each positive source independently, while many approaches - including our humans' manual method - aim at using their synergy. In the next section, we detail another objective to fine-tune these embeddings, inspired by the inherent structure of multihop reasoning.

- After retrieving the correct sources, potentially with some noise, we want to leverage the same embeddings to generate our answer. Indeed, if we have managed to capture some links between positive sources during retrieval, the latter can be leveraged for generation. We explore the conditioning of a small LLM by such multimodal universal embeddings and conclude with respect of building a unified and more performant question answering agent.

##### 4.1.2. TOWARDS A MULTIHOP CONTRASTIVE LOSS

The classic contrastive loss to distinguish a useful source from the others with respect to a specific query is:

$$L_{base} = -\log \frac{e^{f(q, d^+)/\tau}}{e^{f(q, d^+)/\tau} + \sum_{d^- \in \mathcal{D}^-} e^{f(q, d^-)/\tau}} \quad (1)$$

where  $q$  is the query,  $d^+$  a single positive source for  $q$ , and  $d^- \in \mathcal{D}^-$  refer to some negative sources for  $q$  (in-batch or not, cf. below). Here,  $\tau$  is a parameter that controls the temperature scaling after computing the inner product denoted by  $f$  between the embeddings of the query and sources, which all lie in the same universal space of dimension 512.

The results and metrics of (Liu et al., 2023b) are misleading for the challenge posed by (Chang et al., 2022), the latter using the F1 score. From the candidate sources ordered by decreasing cosine-similarity score with the query, how do we chose the decision threshold in practice? A preliminary error analysis shows that the best maximum F1 score reachable by imposing a single threshold is not very good. In the hope of having a better decision boundary for source selection with the same overall contrastive learning, we need to consider the positive sources together, either jointly (2) by opposition with the rest of the cofounders, either sequentially (3).

$$L_{joint} = -\log \frac{e^{f(q, d_1^+)/\tau} + e^{f(q, d_2^+)/\tau}}{e^{f(q, d_1^+)/\tau} + e^{f(q, d_2^+)/\tau} + \sum_{d^- \in \mathcal{D}^-} e^{f(q, d^-)/\tau}} \quad (2)$$

$$L_{hop} = -\log \frac{e^{f(q, d_2^+ | d_1^+)/\tau}}{e^{f(q, d_1^+ | d_2^+)/\tau} + \sum_{d^- \in \mathcal{D}^-} e^{f(q, d^-)/\tau}} \quad (3)$$

In (2), we simply consider the positive sources (2 in this case, which is the case for most queries in WebQA) together. We do not hope this to solve our issue since this objective

alone does not guarantee a similar proximity of all positive sources with the query and thus a single thresholding decision criteria. However, after a first round of training each positive source independently (objective 1, this could help to align the similarities of sources linked by the same question.

We are more excited about the multi-hop contrastive objective in 3, since it is intuitive and yields to a potential sequential multi-hop selection of sources with a single threshold. This objective is in fact similar to (Yang et al., 2023b), but here we operate in the embedding space and not in natural language: the concatenation of the query  $q$  and a first positive retrieved source  $d_1^+$  is a simple addition, similar to how we enhance the visual encoding of an image by its verbalization encoding. Here, the second positive source to retrieve  $d_2^+$  is symmetric with  $d_1^+$ , and we will enforce both configurations during training for consistency. To be clear,

$$f(q, d_1^+ | d_2^+) = (\text{Embed}(q) + \text{Embed}(d_1^+))^T \text{Embed}(d_2^+) \quad (4)$$

*Note:* all the previous equations are adaptable for any number of positive sources.

#### 4.1.3. TRAINING PROCEDURE

The in-batch positive and negative sources for a query in WebQA are not representative of all types of potential sources since they were chosen to all have a similarity with the question. That is why we split the contrastive training in two parts, inspired by (Liu et al., 2023b).

1. We perform in-batch contrastive learning with the base loss function 1. This gives a good first checkpoint for the dual encoders.  
Then, for each query, we defined *hard negatives*, i.e. irrelevant sources that are close to the embedding of the query after the first contrastive pretraining. We will then chose these hard negatives as distractors during the next phase of training, since they are more diverse and not *contrasted* enough by the model at this point.
2. We continue the training with these hard negatives, while making sure we take a similar number of negative sources from each modality for each query/source to be contrasted, e.g.  $D^-$  is composed of 4 distractor images and 4 distractor texts. For this phase, we tried different combinations of the loss functions presented above, this will be detailed in the Experimental Setup.

#### 4.1.4. GENERATION FROM UNIVERSAL EMBEDDINGS

**Generation:** In this study, we propose a novel approach that integrates a language model capable of processing both textual and visual information to address the inherent challenges in multi-modal and multi-hop question-answering.

This model provides the capability of taking in input of multiple images for question answering. Our method leverages a pre-trained language model with frozen weights. The key innovation lies in the utilization of a Contrastive Language-Image Pre-training (CLIP) encoder that were previously trained for retrieval for encoding both image and text context as a new CONTEXT token for a LLM and subsequently, a linear layer is employed to transform this context encoding into a format compatible with the language model, resulting in a set of four tokens. This builds upon the idea from Fromage(Koh et al., 2023) on how to take an image as input to an LLM.

#### Model Description

**Generation:** We denote the image input as  $I$  and the corresponding question as  $Q$ . The CLIP encoder is represented by  $\text{CLIP}(I)$ , which encodes the visual content of the image. The linear transformation is denoted as  $W$ , and the language model is represented by  $\text{LM}(Q)$ . The output of our model, representing the probability distribution over possible answers given the input image and question, is denoted as  $P(\text{answer}|I, Q)$ .

The transformation from image encoding to the language model input can be expressed as follows:

$$\text{LM}_{\text{input}} = W \cdot \text{CLIP}(I)$$

The resulting  $\text{LM}_{\text{input}}$  is then fed into the language model to generate the probability distribution over possible answers:

$$P(\text{answer}|I, Q) = \text{LM}(\text{LM}_{\text{input}}, Q)$$

#### Learning

**Generation:** In the context of Question answering, our learning strategy involves formulating the task as the generation of textual tokens conditioned on a context prefix. The context prefix is constructed through the context-to-text mapping layer and is subsequently prepended to the Question. Given a question  $x$  tokenized as  $(s_1, \dots, s_T)$  and its corresponding context  $y$ , the log likelihood  $l(x, y)$  is expressed as:

$$l(x, y) = \sum_{t=1}^T \log p(s_t | \text{CLIP}(y)^T W, s_1, \dots, s_{t-1}; \theta_t, \phi_c)$$

The loss  $L$  is then defined as the negative log likelihood of all samples in a batch of  $N$  Q-A pairs:

$$L = -\frac{1}{N} \sum_{i=1}^N l(x_i, y_i)$$

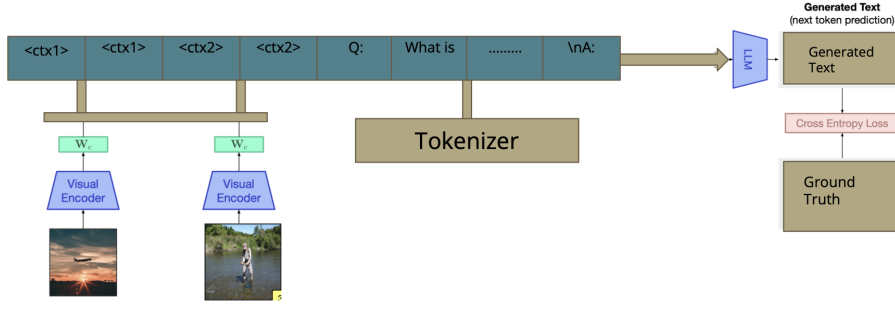


Figure 1. Multimodal Generation Model

## 4.2. Multimodal Question Decomposition Model

In our analysis of the dataset, we observed that about 44 percent of image-based queries and 99 percent of text-based queries need two sources to answer these queries. These questions require the model to understand the structure of the question, correctly retrieve the two sources relevant to the question, derive the information from these sources and aggregate this information. It was seen that the structure of these questions in the WebQA dataset is complex and hard to understand, and would benefit from decomposition and simplification. This is in conjunction with how humans decipher complex data in real life. . Thus, the research idea proposes to break down the question into sub-questions and implement a multi-hop architecture to solve the multi-hop nature of the problem. From the leading submissions of the dataset as well as our error analysis on the baseline models, we see that the performance of the network suffers due to poor generation rather than retrieval. One of the key reasons behind this was identified as the inability of the network to comprehend its retrieved sources correctly with respect to the query. Therefore, we assume that the correct sources have been retrieved and generate answers for the subquestions obtained from the model described below.

### 4.2.1. MODEL DESCRIPTION

A novel model architecture as shown in Figure 2 is proposed, which extends some of the earlier ideas of unimodal text-based question decomposition[] to be used in a multimodal setting. Each of the subquestions is focused on one source for deriving part of the final answer. The methodology can be explained as below:

1. Decompose the question  $Q$  into sub-questions  $Q1$  and  $Q2$
2. For each subquestion  $Q_i$ , obtain partial answers  $A_i$  using multimodal QA model (MQA).

3. Aggregating the initial query  $Q$  with the partial answers  $A_i$  to obtain the final answer with a text QA model (TQA).

### 4.2.2. LEARNING

For decomposition of the question, we use the span-prediction based DecompRC (Min et al., 2019) model which is trained on the multihop unimodal questions of the HotpotQA dataset. As explained by the equations below, A Pointer function  $P$  identifies  $c$  indices in the input sequence,  $S$ . The  $n$  words in the sequence are encoded using BERT where  $h$  is the output dimension of the encoder. A trainable weight matrix  $W$  of of dimension  $h \times c$  is used to compute the pointer score matrix  $Y$  of dimension  $n \times c$  where each element denotes the probability that the  $i$ th word of sequence  $S$  is the  $j$ th pointer index. The model extracts the  $c$  indices that have highest joint probability. For intersection type questions, it identifies two pointer positions to divide the question into sub-questions.

$$U = \text{BERT}(S) \in R^{n \times h}$$

$$Y = \text{softmax}(UW) \in R^{n \times c}$$

$$\text{ind}_1, \dots, \text{ind}_c = \text{argmax}_{i_1 < \dots < i_c} \prod_{j=1}^c P(i_j = \text{ind}_j)$$

## 4.3. Exploiting the emergence properties of Question-Answering tasks

When performing question-answering using information from a set of provided sources, we can argue that the answer usually emerges from the interactions between the query and the sources. However, the current training objectives, such as maximum likelihood training for answer generation don't explicitly exploit the fact that the final answer arises from such interactions between sources and the query.



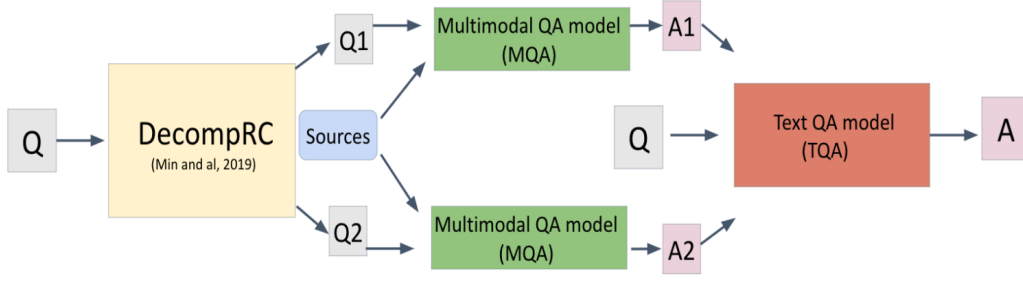


Figure 2. Multimodal Question Decomposition model

Drawing from the idea of Partial Information Decomposition (Liang et al., 2023; Bertschinger et al., 2014), we hypothesize that using training objectives which aim to maximize the synergy between sources and query should be more ideal for question-answering tasks. In this work, we explore whether altering the usual maximum likelihood training objective slightly, with the aim of preventing the model from over-relying on either sources or query for answer generation, helps with overall performance.

#### 4.3.1. MODEL DESCRIPTION

The usual training objectives for question-answering models can be described as:

$$\theta^* = \operatorname{argmax}_{\theta} P_{\theta}(a | s_1, \dots, s_n, q)$$

Where,  $a$  is the actual answer to the query  $q$ ,  $s_1, \dots, s_n$  are the sources to use information from. We add two additional terms to the training objective, both of which penalise the model for over-relying on either source or query. Our training objective is defined as:

$$\theta^* = \operatorname{argmax}_{\theta} [P_{\theta}(a | s_1, \dots, s_n, q) - \alpha_s P_{\theta}(a | s'_1, \dots, s'_m, q) - \alpha_q P_{\theta}(a | s_1, \dots, s_n, q')]$$

Where  $s'_1, \dots, s'_m$  are sources corresponding to any random query, and similarly  $q'$  is any random query in the data. The idea behind this objective is that we don't want model to be able to generate the answer if we provide with the correct query along with random sources and vice-versa.

#### 4.3.2. LEARNING

The overall learning paradigm for our objective is described in Figure 3. We perturb a fraction of examples by either changing the sources or the query, and pass them through our backbone model which gives us the loss on that particular (query, source, answer) tuple. If the example fed to network was unperturbed we perform usual training minimizing the loss. However, if the example was perturbed we multiply the loss by a small negative constants,  $-\beta_s$  and  $-\beta_q$ , to

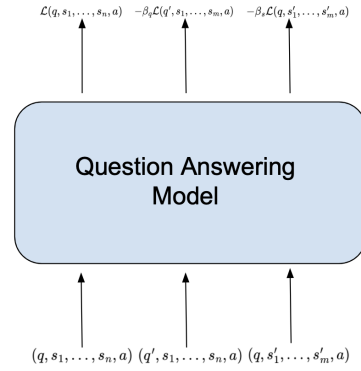


Figure 3. This figure describes the learning paradigm for our method.

decrease the probability of the true answer being outputted when either the query or source is changed.

Overall this method is agnostic of the model used and can be used with any question-answering model.

## 5. Experimental Setup

### 5.1. Universal retrieval and generation using Univl-DR

#### 5.1.1. CONTRASTIVE LEARNING

We use CLIP model by OpenAI, ViT 32GB. This is the base model used (Liu et al., 2023b). We trained it with hyperparameters:  $batch\_size = 32$ ,  $learning\_rate = 5e - 6$ , and a *cosine\_lr* scheduler.

The first checkpoint of the model after in-batch contrastive learning was frozen and we took from there to refine the representations with our proposed learning objectives. For each round of training from this checkpoint, we used a reduced version of the training and validation dataset, with 4 hard-negative sources from each modality - which we shuffled.

*Note:* Due to the RAM necessary for this model, we could not use it on a AWS instance, and were limited by local resources for training - hence we could not perform more than a few epochs of training on a smaller subset of the training data (5000 queries). We are convinced that with more computing power we can demonstrate more substantiate improvements - see Results section.

### 5.1.2. EXPERIMENTS AND CONFIGURATIONS FOR THE RETRIEVAL

We evaluate the retrieval with the F1 metric (and not the ones from (Liu et al., 2023b) which do not suit for this challenge) - on 100 queries from the test set of WebQA (due to time constraints). We compare their original baseline with our retrained models with different configurations:

- **Config 1:** We retrained our model with 1, considering all positive sources independently (with the same negative sources each time, which we could enhance by data augmentation in the future). The idea is that the query embedding is close to each positive source, independently. This is already an improvement with respect to the checkpoint provided in (Liu et al., 2023b), since they only considered one (random) positive source per query during their training.
- **Config 2:** Since 2 is unlikely to bring improvements in itself, we tried a custom combination of the loss functions, in the hope that the combined objectives will bring more stability and balance between the prediction of the positive sources:

$$L = 0.7 * L_{base}(1) + 0.1 * L_{joint}(2) + 0.2 * L_{hop}(3)$$

- **Config 3:** We used the multihop objective 3 after a first round of classic contrastive pretraining. We made sure to hop from source 1 to source 2 and source 2 to source 1 during training.

The F1 score was computed with the method of setting a binary threshold for source selection given all cosine-similarity scores and true labels for sources.

### 5.1.3. GENERATION FROM UNIVERSAL EMBEDDINGS

For the part of generation we use 1.3B parameter facebook OPT LLM. Apart from the UNIVL-DR embeddings we also tried out detectron 2 object detection features with 8 visual tokens to embed 2 images. Training of the generation part is done on 10,000 samples from train dataset. We use AdamW as the optimizer with a learning rate of 0.0009 and 0.9 and 0.99 values for beta1 and beta2. The training was done 30 epochs on a Single 24GB A10G Nvidia GPU which took 6 hours for object detection features and 3 hours

for UniVLDR features. The feature calculation in both the process is done independently in separate process to save GPU memory.

### 5.1.4. MULTIMODAL BASELINE MODELS

Across different research ideas, we use the model Image2LLM VQA as a baseline for generation. This baseline gives us insight about how LLMs can be used for VQA. This model is optimized for single Image VQA so we alter the model to use it for multi-image VQA. Image2LLM VQA enhances the performance of a Language Model (LLM) for Visual Question Answering (VQA) by incorporating image content through the generation of synthetic question-answer pairs and question-relevant image captions. For the WebQA dataset we utilize this model Image-based queries answer generation. Since questions in the WebQA dataset require comparison between multiple images and also sometimes reasoning over multiple images the baseline didn't perform well.

## 5.2. Multimodal Question Decomposition Model

Our research aims at decomposing multimodal multihop questions into subquestions which can be used to derive information from one source at a time. We focus on 'Intersection' type of questions [decompRC] present in the WebQA dataset that require to extract their answers from two text based or two image based sources.

### 5.2.1. EXPERIMENTAL METHODOLOGY

Our first set of experiments are conducted to derive subquestions for these subquestions and table 1 gives details of example text and image based queries decomposed in a zero-shot approach on pre-trained DecompRC model trained on the HotpotQA dataset which is a unimodal text based multihop dataset. While DecompRC has a reading comprehension pipeline for generating the results on the subquestions, we only utilize the subquestion generation module for our usecase. The WebQA queries were converted to SQUAD format for compatibility with the DecompRC pipeline.

Once the subquestions are obtained from the decompRC model, we benchmark our results on several generative pipelines. The Baseline VLP based model which is provided by the WebQA authors was used to identify a subset of Intersection questions which gave incorrect responses, i.e. failure cases and we therefore, start with a near zero accuracy and see the impact of decomposition of the question into simpler subquestions on the answer generation.

Lastly, we take a detailed look at some failure cases and use our degenerative model to do root-cause analysis on the source of error.

### 5.2.2. MULTIMODAL BASELINE MODELS

We demonstrate the ability of our model to leverage several pre-trained multimodal and text-based QA models in our plug-and-play architecture. The MQAs and TQA described in our pipeline can be replaced by any readily available pre-trained and fine-tuned QA model to evaluate their performance. Our initial error analysis was done on the VLP based baseline (Chang et al., 2022). We run experiments on VLP, FROMAGE and Mini-GPT4 and compare the impact of Question decomposition.

### 5.3. Exploiting the emergence properties of Question-Answering tasks

To validate the feasibility of our idea, we compare the performance of model trained on WebQA with maximum likelihood training objective for answer generation against our training objective. Also, to see if each component of added loss term is helpful for performance we do studies with using only one of the added loss terms instead of both of them.

Due to computational limitations, we only use nearly 20% of the overall training data (each data point is sampled with 20% probability). This sampled dataset is kept same across all models. To evaluate our models, we use the whole validation dataset for comparison, as true answer labels aren't publicly available for the test dataset.

#### 5.3.1. MULTIMODAL BASELINE MODELS

As this method requires a backbone question-answering model, we chose to use SKURG (Yang et al., 2023a) as this was the best performing model on the leaderboard which was publicly available.

This architecture does a unified retrieval and generation by the help of structured knowledge. As text encoder and image encoder usually generate representations independently, they link these two together by generating a knowledge graph linking sources based on shared entities between them.

#### 5.3.2. EXPERIMENTAL METHODOLOGY

We first fine tune BART-base (Lewis et al., 2020) and OFA-base (Wang et al., 2022) model on SQuAD 2.0 (Rajpurkar et al., 2018), in-line with the original training procedure performed in training of SKURG. We then use these fine-tuned models for training on WebQA.

We train all the models for 8000 training steps with an effective batch size of 4 and take the best performing model on validation set for comparison. For our method, we provide a true example with 80% probability and a perturbed example with 20%. When using both additional terms, each kind of

perturbation occurs with an equal chance (10% overall).

AdamW (Loshchilov & Hutter, 2019) was our choice of optimizer as it achieved both quick and better convergences than SGD. The learning rate was same across all experiments and had the value  $1e-5$ . For our experiments, both  $\beta_q$  and  $\beta_s$  were set to be  $5e-5$ , as models diverged for values above  $1e-4$ .

## 6. Results and Discussion

### 6.1. Universal retrieval and generation using UNIVL-DR

The results can be found in Table 3.

#### 6.1.1. RETRIEVAL

- The multi-hop helps a lot for retrieval, allowing the F1 score to be high, with a recall of 1 and a precision of 0.78 for source selection across these 100 queries. By setting a single threshold and accepting the first source with higher cosine similarity score above this threshold and doing this iteratively by enhancing the query vector with the retrieved sources, the retrieval is more costly (iterative) but more efficient.
- The other Config 1 and 2 did not really improve the F1 score, compared to the baseline. Config 1 indeed still optimizes over the positive sources independently. We still guess that with more training epochs and data, training over all positive sources will improve the baseline.
- As for Config 2, it is quite unclear if the improvement on F1 score compared to the baseline is really telling of more "stability". However, the great boost given by the introduction of the multi-hop loss gives hope to derive a single-stage objective loss function incorporating contrasting each positive source individually and performing multi-hop retrieval with enhanced queries.

#### 6.1.2. GENERATION

The generation was fine-tuned on features from config 1 and used off the shelf for other 2 configurations. The results improve for config 3 with increase in retrieval score. Fromage finetuning helps in answering logic based questions in the dataset but the model features are not able to capture fine grained image features. Replacing clip embeddings with more dense visual features and more visual tokens can help in this.

### 6.2. Multimodal Question Decomposition Model

- Table 1 shows the decomposition of some text and image based queries using the DecompRC model. We see



Type	Original Question	Sub-question 1	Sub-question 2
Text	What type of body part do both the saphenous nerve and the superior lateral brachial cutaneous nerve belong to ?	What type of body part do both the saphenous nerve belong to ?	What type of body part do the superior lateral brachial cutaneous nerve belong to ?
Text	from what hill can you see both the basilica julia and the temple of romulus?	from what hill can you see both the basilica julia ?	from what hill can you see the temple of romulus?
Image	what shape are the headlights on both the 1952 kaiser manhattan and the lotus xi lemans?	what shape are the headlights on both the 1952 kaiser manhattan ?	what shape are the headlights on the lotus xi lemans?
Image	are there steps at the entrance to the hood museum of art and budynek biblioteki?	are there steps at the entrance to the hood museum of art ?	are there steps at budynek biblioteki?

Table 1. Subquestions Generated using DecompRC

that the model does a decent task of decomposing the ‘Intersection’ type of multihop questions. We notice the presence of some skew words such as ‘both’ and ‘share’, which are found to not impact the outcome of the sub-query significantly however, it can be improved upon.

- Table 2 displays the results from the entire pipeline using the VLP baseline model proposed by the authors including the *original question*, *expected answer* and the *original answer* obtained from the model followed by the *subquestions* and the *responses* to each of them and in the end, the *final answer* generated by combining the responses to the subquestions in a text-based decoder model. For these results, the text-decoder is also VLP. However, the model is flexible and one or more of the components can be replaced with more efficient multimodal or text based QA models.
- In Table 2, we see that some queries that were failing earlier have been successfully resolved by the decomposition of the question. For example, the first query is a text-based query ‘*What biological classification do pantherinae and the felinae have?*’ for which the original model was unable to form a correct response however, after it was decomposed into the two subquestion ‘*what biological classification do pantherinae?*’

and ‘*what biological classification the felinae have?*’, the correct response ‘*subfamily*’ was generated.

- As described in Table 4, we experimented with VLP, FROMAGE and Mini-GPT4. For VLP, we saw an improvement in the accuracy from 0.052 to 0.263 by the use of question decompositions. For FROMAGE, the accuracy jumped from 0.315 to 0.473. While the fluency dropped in the case of VLP baseline, it improved for FROMAGE. One suspected reason for decline in fluency could be that the decomposition of questions followed by later reconstruction of the answers could be effecting the natural language structure.
- The last two rows of Table 2 give two failure cases, where the answers didn’t improve after the question decomposition. However, these provide an apt example of the enhanced ‘**Interpretability**’ of the model proposed by us. In the question ‘*What color do the stalks of the Speise Morchel and the Amanita phalloides share ?*’, we see that the desired answer is ‘*white*’ but the model outputs ‘*yellow*’. Initially, it is hard to point to the source of error. However, after looking at the response to the subquestions, we can clearly understand that the model is confusing the white color as yellow in both the images. On replacing VLP with Mini-GPT4, it was found that the color of stalks is correctly identi-

fied and the text based decoder successfully combines the ‘sub-answers’ to give the final answer as ‘white’. Similarly, in the last query based on the two paintings ‘Peace and Plenty’ and ‘Claude’s landscape’ we find that the visual comprehension is unable to identify the animals in the image correctly and the model seems to be hallucinating names of animals. This establishes the fact that the question decomposition can be used as a powerful tool to analyze errors in multi-source models and compare the performance of different models

### 6.3. Exploiting the emergence properties of Question-Answering tasks

To evaluate the retrieval capabilities of model, we calculate the precision and recall for the retrieved passages. Whereas, to evaluate the quality of answers generated by the model, we use two metrics defined in the original paper of WebQA, namely fluency and accuracy. Mathematically, these metrics are defined as:

$$\text{FL}(c, R) = \max \left\{ \min \left( 1, \frac{\text{BARTS core}(r, c)}{\text{BARTS core}(r, r)} \right) \right\}_{r \in R}$$

$$\text{Acc}(c, K) = \begin{cases} \text{if } qc \in [\text{color, shape, number, Y/N}] : \\ F1(c \cap D_{qc}, K \cap D_{qc}) \\ \text{otherwise :} \\ RE(c, K) \end{cases}$$

The final results of our experiments are summarized in Table 5. From the numbers we can see that our prescribed method doesn’t lead to any improvements in performance, instead causes a slight dip in performance. This can be attributed to many reasons such as: (a) maximum likelihood training generalizes and is sufficient to identify that interactions which lead to emergence are most important for this task, (b) this objective isn’t mathematically grounded to force model to maximize the synergy between query and sources, and naive fixes like this aren’t good enough and need for careful reasoning when formulating the problem, and (c) we didn’t perform an active search on hyperparameter space and the current choice of hyperparameters isn’t ideal. We had to choose very small values for  $\beta_s$  and  $\beta_q$  and some components of loss term in SKURG diverged quickly even though other terms were behaving nicely.

So, a possible avenue of future work could to explore this training objective with other models, as this method is model agnostic, which allow for larger values of these hyperparameters. Another possible direction could be to find better formulation for optimization problem which are more in line with the Partial Information Decomposition framework mathematically.

## 7. Conclusion and Future Directions

Our different research ideas explored different complementary directions, pointing towards more robust, multimodally-grounded and scalable end-to-end question-answering assistants.

We found that the current contrastive learning objective can be refined for a more efficient source selection by introducing a multi-hop contrastive loss (3), which makes the source selection iterative with a single threshold. With these universal embeddings, we experimented that generation of answers can be done, since a simple mapping from this space to the token input space of a small LLM yields coherent answers. We believe that by pushing our retraining of the embeddings further and exploring larger generative models or simpler decoders from these latents, a scalable and unified retrieval-into-generation pipeline is within reach. In the short term, it will be interesting to see if the loss from the LLM output can be used to alter the embeddings to improve results for both retrieval and generation.

The Question decomposition model showed improvement on several prior failure cases, both in text and visual modalities and while using different baselines i.e. VLP and FRO-MAGe. It provides high value in the form of interpretability of intermediate results in multihop architectures and helps us understand the chain of thought reasoning of the model. For future directions, we can look into abridging long questions which work on single sources but have complex language. Preliminary experiments on this use case have shown positive results.

We also experimented with different optimisation formulation for this task and from our results we can see that there’s a need for proper mathematically grounded formulation if we want to use such methods which focus on different type of interactions across modalities.

## 8. Contributions

### 8.1. Manh-Bao

Worked on the idea of improving and leveraging universal embeddings for retrieval and generation. Proposed and experimented new contrastive objectives as well as training procedures. Was helped by Avi for the generation part from the obtained embeddings.

### 8.2. Shruti

Worked on the idea and implementation of Multimodal Question Decomposition Model to decompose the questions using the DecompRC model and then ran experiments on the VLP Baseline model, with and without decompositions. Identified the data subset for decomposition experiments. Ran Mini-GPT4 based experiments on this model as well.

### 8.3. Sushil

Worked on the idea of penalising model for predicting the generated answer when we corrupt the sources or the query provided to the model. This involved setting up the backbone model and running different experiments to evaluate if the method is helpful for the task or not.

### 8.4. Avi

Created generator models used with 2 ideas and worked with shruti on Question Decomposition model. Created features using 2 different image encoders. Setup WebQA baseline for training. Performed inference on fromage for multiple experiments. Inference using MiniGPT and ChatGPT for Shruti's idea.

## References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning, 2022.
- Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4): 2161–2183, apr 2014. doi: 10.3390/e16042161. URL <https://doi.org/10.3390/e16042161>.
- Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. Webqa: Multihop and multimodal qa, 2022.
- Chen, J., ting Lin, S., and Durrett, G. Multi-hop question answering via reasoning chains, 2021.
- Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal inputs and outputs, 2023.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetraault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703>.
- Liang, P. P., Cheng, Y., Fan, X., Ling, C. K., Nie, S., Chen, R., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., and Morency, L.-P. Quantifying & modeling multimodal interactions: An information decomposition framework, 2023.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023a.
- Liu, Z., Xiong, C., Lv, Y., Liu, Z., and Yu, G. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval, 2023b.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.
- Malon, C. and Bai, B. Generating followup questions for interpretable multi-hop question answering, 2020.
- Mavi, V., Jangra, A., and Jatowt, A. A survey on multi-hop question answering and generation, 2022.
- Min, S., Zhong, V., Zettlemoyer, L., and Hajishirzi, H. Multi-hop reading comprehension through question decomposition and rescoring, 2019.
- Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad, 2018.
- Sun, H., Cohen, W. W., and Salakhutdinov, R. Iterative hierarchical attention for answering complex questions over long documents, 2021.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.
- Xiong, W., Li, X. L., Iyer, S., Du, J., Lewis, P., Wang, W. Y., Mehdad, Y., tau Yih, W., Riedel, S., Kiela, D., and Oğuz, B. Answering complex open-domain questions with multi-hop dense retrieval, 2021.
- Yang, Q., Chen, Q., Wang, W., Hu, B., and Zhang, M. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, oct 2023a. doi: 10.1145/3581783.3611964. URL <https://doi.org/10.1145/3581783.3611964>.
- Yang, S., Wu, A., Wu, X., Xiao, L., Ma, T., Jin, C., and He, L. Progressive evidence refinement for open-domain multimodal retrieval question answering, 2023b.
- Yu, B., Fu, C., Yu, H., Huang, F., and Li, Y. Unified language representation for question answering over text, tables, and images, 2023.
- Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.








Type	Initial query	Expected Answer	Answer Generated by VLP Baseline	Subquestion 1 and Answer	Subquestion 2 and Answer	Final Answer
Text	What biological classification do pantherinae and the felinae have?	They are both an example of subfamily, under the larger name of Felidae.	They are both category	what biological classification do pantherinae? <b>Ans:</b> subfamily	what biological classification the felinae have? <b>Ans:</b> subfamily	subfamily
Text	What type of rails do the West Rail line in Denver and the Mass Transit Railway in Hong Kong have in common ?	The West Rail line and the Mass Transit Railway both use light rails	They are both tube - rail systems	what type of rails do the west rail line in denver ? <b>Ans:</b> Light rails	what type of rails do the the mass transit railway in hong kong have in common? <b>Ans:</b> light rails	Light rails
Text	Which feature is shared by Male Crickets and Micropterigidae ?	Teeth	Male insects are both insects .	Which feature is shared by Male Crickets? <b>Ans:</b> Male crickets are known for their chirp ( which only have ridges or a curl vein ) which bears down to 300 teeth	Which feature is shared by Micropterigidae? <b>Ans:</b> The male wings of the Micropterigidae are a thick rib and a curved tail	Male crickets and Micropterigidae have ridges or a curved vein
Image	Is the hair part lined up with the person ' s nose in both Her Know and Joseph Two Bulls ? 	Yes, the hair part is lined up with the person's nose in both Her Know and Joseph Two Bulls. 	No , the hair part is not lined up with the person ' s nose in both Her Know and Joseph Two Bulls	is the hair part lined up with the person's nose in both her know ? <b>Ans:</b> Yes , the hair part is lined up with the person ' s nose in Her Know	is the hair part lined up with the person's nose in joseph two bulls? <b>Ans:</b> The hair part is lined up with the person ' s nose in Joseph Two Bulls	Yes , the hair part is lined up with the person ' s nose in Her Know and Joseph Two Bulls
Image	What color do the stalks of the Speise Morchel and the Amanita phalloides share ? 	They share the color white. 	The stalks of both are red 	What color do the stalks of the speise morchel ? <b>Ans:</b> The stalks of the Speise Morchel are yellow.	What color do the amanita phalloides share? <b>Ans:</b> The color of the Amanita phalloides is yellow .	The stalks of the Speise Morchel and the Amanita phalloides are yellow .
Image	What animal can be found in both Peace and Plenty and Claude ' s landscape? 	A cow can be found in both Peace and Plenty and Claude's landscape 	A dog can be found in both Peace and Plenty and Claude ' s landscape	What animal can be found in both Peace and Plenty? <b>Ans:</b> Birds can be found in Peace and Plenty	What animal can be found in Claude ' s landscape ? <b>Ans:</b> A dog is found in Claude's landscape	Birds

Table 2. Questions Decomposition and Expected vs Generated Answers

	Accuracy	Fluency	F1-Score
<b>Baseline</b>	-	-	0.65
<b>Config 1</b>	0.512	0.162	0.59
<b>Config 2</b>	0.404	0.149	0.69
<b>Config 3</b>	0.444	0.151	0.87

Table 3. Results for retrieval and generation using FROMAGe with fine-tuned UNivL-DR embeddings



Model	Accuracy	Fluency
VLP (without decomposition)	0.052	0.3
VLP (with decomposition)	0.263	0.21
FROMAGe (without decomposition - Univl-dr features)	0.315	0.15
FROMAGe (with decomposition - Univl-dr features)	0.473	0.18
FROMAGe (with decomposition - visual features)	0.394	0.19
MiniGPT-4 (visual queries) + ChatGPT (text-based queries)	0.552	0.16

Table 4. Fluency and Accuracy of different MQA and TQA models on a subset of intersection queries from validation dataset

Model	Precision	Recall	Accuracy	Fluency	Acc*Fluency
Baseline	0.8153	0.6593	0.5372	0.4282	0.3134
Only queries perturbed	0.8323	0.6334	0.5207	0.4218	0.3089
Only sources perturbed	0.8094	0.6407	0.5174	0.4204	0.3049
Both queries and sources perturbed	0.8098	0.6380	0.5196	0.4176	0.3041

Table 5. Results for changing the optimisation problem to account for emergence properties of interactions between sources and query. Higher is better for all metrics